EMBOSS User's Guide Practical Bioinformatics

Mr Peter Rice
EMBL European Bioinformatics Institute

Dr Alan Bleasby
EMBL European Bioinformatics Institute

Dr Jon Ison
EMBL European Bioinformatics Institute

with contributions from
Lisa Mullan
Guy Bottu

CAMBRIDGE
UNIVERSITY PRESS

Contents

Acknowledgements *page* xvi

Preface xix

Conventions xxii

Welcome to the *EMBOSS User's Guide* xxv
Summary xxv
Chapter 1. Background to EMBOSS xxv
Chapter 2. Basic setup and maintenance xxv
Chapter 3. Getting started xxv
Chapter 4. Tutorial xxv
Chapter 5. File formats xxv
Chapter 6. The EMBOSS command line xxvi
Chapter 7. Interfaces xxvi
Chapter 8. Using EMBOSS under **wEMBOSS** xxvi
Chapter 9. Using EMBOSS under **Jemboss** xxvi
Appendix A. File format reference xxvi
Appendix B. Application reference xxvii
Appendix C. Command-line qualifier reference xxvii
Appendix D. Resources xxvii

1 **Background to EMBOSS** 1
1.1 History 1
1.2 EMBOSS developers 2
1.3 Key features 3
1.3.1 General features 3
1.3.2 Features for users of EMBOSS 5

2 **Basic setup and maintenance** 7
2.1 Supported platforms 7
2.2 Hardware requirements 8
2.3 Software requirements 8
2.3.1 GNU tools 8
2.3.2 EMBOSS dependencies 8
2.3.3 EMBASSY dependencies 8
2.4 Software releases 9
2.4.1 Stable releases 9
2.4.2 Developer's (CVS) release 11
2.5 Downloading the stable release 11
2.5.1 Downloading via the EMBOSS website 11
2.5.2 Downloading via anonymous FTP 12
2.6 Package structure 14

2.6.1 Major components 14
2.6.2 Sub-components 14
2.6.3 Differences between CVS and stable versions 15
2.7 Installation 16
2.7.1 Overview of the installation process 16
2.7.2 Configuration 16
2.7.3 Compilation 17
2.7.4 Setting your PATH 18
2.7.5 Testing all is well 18
2.7.6 Database setup 19
2.7.7 Installing EMBASSY packages 20
2.8 Maintenance 22
2.8.1 Using CVS to update 22
2.8.2 Bug-fix replacement files 23
2.8.3 Patch files 24
2.8.4 Automated installation of EMBOSS and EMBASSY 25
2.8.5 Automated database updating 26

3 **Getting started** 28
3.1 Application documentation 28
3.1.1 Online documentation 28
3.1.2 AJAX command definition (ACD) language 29
3.1.3 Interfaces 29
3.2 Navigating the application documentation 29
3.2.1 Navigating the tabular documentation 29
3.2.2 Sections in the application documentation 29
3.3 How to contribute 30
3.3.1 EMBOSS coordination meetings 31
3.3.2 Collaborations 31
3.4 Project mailing lists 31
3.4.1 User mailing list 31
3.4.2 Developer mailing list 32
3.4.3 Announcements mailing list 32
3.4.4 Mail archives 32
3.5 How to get help 33
3.5.1 EMBOSS documentation 33
3.5.2 EMBOSS frequently asked questions 33
3.5.3 Asking for help 33
3.5.4 Suggesting new features and applications 34
3.6 Reporting bugs and problems 35
3.6.1 Where to send a bug report 35
3.6.2 Before you send a bug report 35
3.6.3 How to write a bug report 35
3.7 EMBOSS training 35
3.7.1 EMBOSS tutorial 36
3.7.2 EMBOSS developer's course 36
3.7.3 EMBOSS workshops 36

4 **EMBOSS user tutorial** 37
4.1 How this tutorial is organised 37
4.2 **wossname**: a first EMBOSS application 37
4.2.1 Exercise: `wossname` 37
4.3 Working with sequences 38
4.4 Retrieving sequences from databases 39
4.4.1 Exercise: `showdb` 39
4.5 `seqret` 40
4.5.1 Exercise: `seqret` 40
4.6 Reading sequences from files 41
4.7 `infoseq` 41
4.8 Sequence annotation 42
4.9 Using multiple sequences 44
4.10 Listfiles 44
4.11 Pairwise sequence alignment 46
4.12 Dotplots 47
4.12.1 Exercise: making a dotplot 47
4.12.2 Exercise: examining dotplot parameters 48
4.13 Global alignment 48
4.13.1 Exercise: `needle` 49
4.14 Local alignment 50
4.14.1 Exercise: `water` 50
4.15 Protein analysis 52
4.16 Identifying the open reading frame (ORF) 52
4.16.1 Exercise: `plotorf` 52
4.16.2 Exercise: `getorf` 52
4.17 Translating the sequence 54
4.17.1 Exercise: `transeq` 54
4.18 USA for partial sequences 55
4.19 Secondary structure prediction 55
4.20 **pepinfo** 56
4.20.1 Exercise: `pepinfo` 56
4.21 Predicting transmembrane regions 56
4.21.1 Exercise: `tmap` 56
4.22 Patterns, profiles and multiple sequence alignment 58
4.23 Pattern matching 59
4.23.1 Exercise: `patmatmotifs` 59
4.24 Report formats 60
4.25 Protein fingerprints 61
4.25.1 Exercise: `pscan` 61
4.26 Multiple sequence analysis 62
4.26.1 Exercise: retrieving a set of sequences 63
4.26.2 Exercise: `emma` 63
4.26.3 Exercise: `prettyplot` 66
4.27 Profiles 67
4.27.1 Exercise: `prophecy` 68
4.27.2 Exercise: `prophet` 69

4.28 Conclusion 69
4.28.1 Exercise: `tfm` 69

5 **File formats** 70
5.1 Introduction to file formats 70
5.2 Introduction to sequence formats 71
5.2.1 What is a sequence format? 71
5.2.2 Supported sequence formats 72
5.2.3 Contents of a sequence entry 76
5.2.4 Specifying sequences on the command line 80
5.2.5 Applications for basic sequence manipulation 81
5.3 Introduction to feature formats 82
5.3.1 What is a feature? 82
5.3.2 Supported feature formats 82
5.3.3 How are features stored ? 83
5.3.4 Applications for features 84
5.3.5 Specifying features on the command line 84
5.4 Introduction to alignment formats 84
5.4.1 What is an alignment format? 85
5.4.2 Supported alignment formats 86
5.4.3 Contents of an alignment file 87
5.4.4 Specifying alignments on the command line 88
5.4.5 Applications for sequence alignment 89
5.5 Introduction to report formats 91
5.5.1 What is a report format? 91
5.5.2 Supported report formats 91
5.5.3 Inside a report 92
5.5.4 Specifying reports on the command line 94
.5.5 Applications that use reports 94

, **The EMBOSS command line** 96
6.1 Introduction to the EMBOSS command line 96
6.1.1 Finding and running EMBOSS applications 96
6.1.2 Application options 96
6.1.3 Command line styles 102
6.1.4 Environment variables 102
6.2 Specifying values for application options 102
6.2.1 General rules 102
6.2.2 Simple ACD datatypes 103
6.2.3 Input ACD datatypes 106
6.2.4 Output ACD datatypes 111
6.2.5 Selection ACD datatypes 114
6.2.6 Graphics ACD datatypes 116
6.3 Global command line qualifiers 116
6.3.1 Introduction 116
6.3.2 Description of global qualifiers 117
6.3.3 Global qualifiers and environment variables 121

6.4 Datatype-specific command line qualifiers 123
6.4.1 Introduction 123
6.4.2 Sequences 123
6.4.3 Sequence features 132
6.4.4 Sequence alignments 137
6.4.5 General input 139
6.4.6 Patterns 139
6.4.7 General output 140
6.4.8 Application report output 141
6.5 Graphical output 142
6.5.1 Description of qualifiers 142
6.6 The Uniform Sequence Address (USA) 143
6.6.1 Introduction 143
6.6.2 USA syntax 144
6.6.3 Specifying the format 146
6.6.4 Specifying a database 147
6.6.5 Specifying a sequence file 149
6.6.6 Specifying a listfile 151
6.6.7 Specifying a sequence 'as is' 152
6.6.8 Applications 152
6.6.9 Specifying search fields 153
6.6.10 USA summary 157
6.7 The Uniform Feature Object (UFO) 159

7 **Interfaces** 160
7.1 Introduction 160
7.2 Command line interfaces 160
7.3 Types of EMBOSS interfaces 161
7.4 Web interfaces 161
7.4.1 **wEMBOSS** 162
7.4.2 **WebLab** 162
7.4.3 **EMBOSS Explorer** 162
7.4.4 **W2H** 162
7.4.5 **SRSWWW** 163
7.4.6 **BioNavigator** 163
7.4.7 **Spinet** 163
7.5 Graphical user interfaces (GUIs) 163
7.5.1 **Jemboss** 164
7.5.2 **Staden** 165
7.5.3 **CoLiMate** 165
7.5.4 **Kaptain** 165
7.5.5 **kemboss** 165
7.5.6 **Geneious** 166
7.6 Workflow interfaces 166
7.6.1 **Taverna** 167
7.6.2 **metalife** 167
7.6.3 **Pipeline Pilot** 167

7.6.4 **BioWBI** and **WsBAW** 167
7.6.5 **G-Pipe** 168
7.6.6 **Mobyle** 168
7.7 Other interfaces 168
7.7.1 **Utopia** 169
7.7.2 **emnu** 169
7.7.3 **MolTalk** 169
7.8 Selecting an interface 169

8 **Using EMBOSS under wEMBOSS** 170
8.1 Introduction 170
8.2 Managing projects and files 171
8.3 Project management 172
8.3.1 Running programs 172
8.4 Programs 175
8.4.1 Handling input and output 175
8.5 Plug-ins and applets 177
8.6 Bugs and fixes 178
8.7 **wEMBOSS** tutorial 178
8.7.1 Exercise: Starting up **wEMBOSS**, creating a 'project', running a program 179
8.7.2 Exercise: Accessing 'public' databanks, using the sequence selectors, managing graphical output 179
8.7.3 Exercise: Running a program on multiple sequences, using the output of one program as input of another, using plug-ins and applets 180

9 **Using EMBOSS under Jemboss** 183
9.1 Diving in at the deep end 183
9.2 Getting started 185
9.2.1 Software requirements 185
9.2.2 Microsoft desktop 186
9.2.3 Apple Macintosh 186
9.2.4 UNIX platform 186
9.2.5 Local installation 186
9.2.6 Remote installation 187
9.2.7 **Jemboss** session 187
9.2.8 Session-specific information 187
9.2.9 The **Jemboss** windows 187
9.2.10 Settings 188
9.2.11 Proxies 189
9.2.12 Servers 189
9.3 File management 189
9.3.1 Local file management 189
9.3.2 Home directory 189
9.3.3 Working directory 190
9.3.4 Move up a directory 190
9.3.5 Creating data files in **Jemboss** 190

9.3.6 File manipulation 191
9.3.7 New folder creation 191
9.3.8 Re-locating files 192
9.3.9 Rename 192
9.3.10 Delete 192
9.3.11 De-select all 192
9.3.12 Refresh 193
9.3.13 Open with 193
9.3.14 Remote file management 193
9.3.15 EMBOSS results 193
9.3.16 Moving data between file managers 194
9.4 Data analysis 194
9.4.1 Program selection 194
9.4.2 Program categories 194
9.4.3 Favourites 194
9.4.4 Alphabetical program list 195
9.4.5 Go To box 195
9.4.6 Input section 195
9.4.7 File input 195
9.4.8 Input sequence options 196
9.4.9 Databases available 197
9.4.10 Sequence format 197
9.4.11 Begin/end 197
9.4.12 Reverse complement 197
9.4.13 Nucleotide/protein 197
9.4.14 Upper/lower-case 197
9.4.15 UFO features 197
9.4.16 Load sequence attributes 198
9.4.17 Parameter selection 198
9.4.18 Output section 198
9.4.19 Output sequence options 199
9.4.20 Sequence format 199
9.4.21 Filename extension 199
9.4.22 Base filename 199
9.4.23 Features format 199
9.4.24 Features filename 200
9.4.25 Sequence format 200
9.4.26 Graphical format 200
9.4.27 PNG graphics 200
9.4.28 **Jemboss** graphics 200
9.4.29 Graph options 201
9.4.30 Main title 201
9.4.31 Axis number format 201
9.4.32 Ticks 201
9.4.33 Axis labels 201
9.4.34 Graph formatting 201
9.4.35 Saving **Jemboss** graphics 202

9.4.36 Advanced parameter selection 202
9.4.37 Program run options 202
9.4.38 Interactive mode 202
9.4.39 Batch mode 202
9.5 Saving results 203
9.5.1 Saving locally 203
9.5.2 Saved results: interactive mode 203
9.5.3 Saved results: batch mode 204
9.5.4 Saving remotely 204
9.5.5 Analysis run autosave 205
9.5.6 Local autosave 205
9.5.7 Remote autosave 205
9.6 Results retrieval 206
9.6.1 Retrieving interactive results 206
9.6.2 Retrieving batch results 206
9.6.3 *Job Manager* 206
9.6.4 *Current Sessions Results* 206
9.6.5 Display results 207
9.6.6 Delete results 207
9.6.7 Refresh icon 207
9.6.8 Retrieving saved results 208
9.7 Customisation 211
9.7.1 Directory location 211
9.7.2 Program selection 212
9.7.3 Input/output options 212
9.7.4 *Job Manager* update frequency 214
9.7.5 Calculate dependencies 214
9.7.6 Proxy and server settings 214
9.8 Utilities 214
9.8.1 Jemboss Alignment Editor (JAE) 214
9.8.2 DNA Editor 220
9.8.3 JALVIEW 223
9.9 Documentation 223
9.9.1 **Jemboss** user guide 223
9.9.2 Application documentation 223
9.9.3 Version number 224
9.9.4 Tooltips 224
9.10 Troubleshooting 224

Appendix A **File format reference** 226
A.1 Supported sequence formats 226
A.1.1 ABI trace 226
A.1.2 ACEDB 226
A.1.3 ASN1 227
A.1.4 Asis 228
A.1.5 Clustal 228
A.1.6 CODATA 228

A.1.7 DAS 229
A.1.8 DASDNA 230
A.1.9 Debug 230
A.1.10 EMBL 231
A.1.11 Experiment (Staden) 233
A.1.12 FASTA 234
A.1.13 FASTA (GCG) 234
A.1.14 FASTA (Pearson) 235
A.1.15 FASTA (with accession) 235
A.1.16 FASTA (database and identifier) 235
A.1.17 FASTA (GI style) 236
A.1.18 FASTA (NCBI style) 236
A.1.19 Fastq 237
A.1.20 Fastq (Illumina) 237
A.1.21 Fastq (Sanger) 237
A.1.22 Fastq (Solexa) 238
A.1.23 Fitch 238
A.1.24 GCG 8, GCG 9.x and 10.x 238
A.1.25 *GenBank* 239
A.1.26 GenPept 241
A.1.27 GFF3 242
A.1.28 GFF2 243
A.1.29 Hennig86 243
A.1.30 Intelligenetics 244
A.1.31 Jackknifer 244
A.1.32 Jackknifer (non-interleaved) 245
A.1.33 MASE 245
A.1.34 MEGA 246
A.1.35 MEGA (non-interleaved) 246
A.1.36 MSF 247
A.1.37 NBRF/PIR 248
A.1.38 NEXUS/PAUP (interleaved) 248
A.1.39 NEXUS/PAUP (non-interleaved) 249
A.1.40 PDB 250
A.1.41 PDB (nucleotide) 252
A.1.42 Pfam/Stockholm 261
A.1.43 PHYLIP (interleaved) 264
A.1.44 PHYLIP (non-interleaved) 264
A.1.45 Raw 265
A.1.46 *RefseqP* 265
A.1.47 SELEX 268
A.1.48 Staden (obsolete) 269
A.1.49 Strider 269
A.1.50 SwissProt 271
A.1.51 Text/Plain 273
A.1.52 Treecon 274
A.2 Supported feature formats 275

A.2.1 DASGFF 275
A.2.2 *EMBL, GenBank, DDBJ* 279
A.2.3 GFF3 280
A.2.4 GFF2 281
A.2.5 PIR/NBRF 282
A.2.6 SwissProt 283
A.3 Supported alignment formats 284
A.3.1 FASTA 284
A.3.2 Markx0 284
A.3.3 Markx1 285
A.3.4 Markx2 286
A.3.5 Markx3 287
A.3.6 Markx10 288
A.3.7 Match 289
A.3.8 MSF 289
A.3.9 Multiple 290
A.3.10 Pair 291
A.3.11 Score 292
A.3.12 Simple 292
A.3.13 SRS 292
A.3.14 SRS Pair 293
A.3.15 TCOFFEE 294
A.3.16 Trace (debugging only) 295
A.4 Supported report formats 295
A.4.1 DAS GFF feature table 295
A.4.2 Dbmotif 297
A.4.3 Debug report format 298
A.4.4 Diffseq 298
A.4.5 *EMBL* feature table 300
A.4.6 FeatTable 300
A.4.7 *GenBank* feature table 301
A.4.8 GFF feature table 301
A.4.9 Listfile 301
A.4.10 Motif 302
A.4.11 Nametable 304
A.4.12 *PIR* feature table 306
A.4.13 Regions 306
A.4.14 SeqTable 307
A.4.15 SRS 308
A.4.16 SRS Simple 310
A.4.17 *SwissProt* feature table 312
A.4.18 Tab-delimited format 312
A.4.19 Table 313
A.4.20 TagSeq 314
A.4.21 Trace feature table (debugging only) 317

Appendix B **Applications and packages reference** 318
B.1 Applications and packages documentation 318
B.1.1 Online documentation 318
B.2 Application groups (release R6) 319
B.3 EMBASSY packages (release R6) 320
B.4 Applications 321
B.4.1 EMBOSS applications (release R6) 321
B.4.2 EMBASSY applications (available alongside EMBOSS release R6) 326
B.4.3 All applications (by group) 329
B.5 GCG to EMBOSS comparison 343

Appendix C **Command line qualifier reference** 361
C.1 Global qualifiers 361
C.2 Datatype-specific qualifiers 361
C.2.1 Sequence input 361
C.2.2 Sequence output 363
C.2.3 Features 364
C.2.4 Alignments 364
C.2.5 Patterns 365
C.2.6 Outputs 365
C.2.7 Reports 368
C.2.8 Graphics 368

Appendix D **Resources** 369
D.1 EMBOSS servers and portals 369
D.1.1 EMBOSS portals 369
D.1.2 EMBOSS servers 369

Index 371